DOCUMENT RESUME

ED 426 101 TM 029 331

AUTHOR Moore, Jenny
TITLE Some Basic Concepts in Meta-Analysis.
PUB DATE 1999-01-00
NOTE 20p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (San Antonio, TX, January 21-23, 1999).
PUB TYPE Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Comparative Analysis; *Effect Size; *Meta Analysis; Regression (Statistics)
IDENTIFIERS Generality

ABSTRACT

The business of science is formulating generalizable insight. No one study, taken singly, establishes the basis for such insight. Meta-analysis, however, can be used to determine if results generalize and to estimate the mean and the variance of effect sizes across studies (J. Hunter and F. Schmidt, 1990). Meta-analysis inquiries treat studies (rather than people) as the units of analysis, and then use regression or other methods to determine the study features that explain or predict variability in detected effects in a given area of inquiry. There are two primary approaches to meta-analysis that are used: the comparison approach and the correlation approach. This paper explains the basic processes of doing a meta-analysis from the mean difference (comparison) approach to determine which of two independent groups or methods scores higher on the variable of interest. (Contains 31 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 426 101

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Jenny Moore

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Some Basic Concepts in Meta-Analysis

Jenny Moore

Texas A&M University

TM029331

Paper presented at the annual meeting of the Southwest Educational Research Association, San Antonio, January 1999

ABSTRACT

The business of science is formulating generalizable insight. No one study, taken singly, establishes the basis for such insight. Meta-analysis, however, can be used to determine what results generalize, and to estimate the mean and the variance of effect sizes across studies (Hunter & Schmidt, 1990). Meta-analysis inquiries treat studies (rather than people) as the unit of analysis, and then employ regression or other methods to determine which study features explain or predict variability in detected effects in a given area of inquiry. There are two primary approaches to meta-analysis that are utilized, the comparison approach and the correlation approach. The present paper is on the basic processes of doing a meta-analysis from the mean difference (comparison) approach to determine which of two independent groups or methods scores higher on the variable of interest.

As Glass (1977) put it, "there are no perfect studies, but it is reasonable to believe that together imperfect studies can converge on a legitimate conclusion" (p. 356). Meta-analysis statistically analyzes research data from a collection of independent studies testing the same hypothesis. The meta-analytic process synthesizes results across many studies. Meta-analysis can be employed to synthesize results across both *substantive studies* (e.g., what if the average effect across studies of variations in class size, or of counseling?) and measurement studies. As regards measurement meta-analyses, both *validity generalization* (Hunter & Schmidt, 1990; Schmidt & Hunter, 1977) and *reliability generalization* (Vacha-Haase, 1998) methods have been articulated.

The desire to generalize across many studies is legitimate because a single statistical significance test does not indicate result importance or insure that results will be generalized across future studies (e.g., Carver, 1978; Cohen, 1994; Daniel, in press; Kirk, 1996; McLean & Ernest, in press; Meehl, 1978; Nix & Barnette, in press; Thompson, 1996, 1998a, 1998b, 1998c, in press-a, in press-b, in press-c). With meta-analysis large amounts of information can be managed, and the statistical process can be considered more objective than the narrative review process that lends itself to personal biases (Hyde, 1986). If results across studies with a shared hypothesis are consistent, then the mean effect size and its standard error will be a reasonable summary of the relation between the independent variable(s) and the dependent variable(s) (Kalain & Raudenbush, 1996). The purpose of this paper is to offer an introduction to meta-analysis for those with little knowledge of the process; Kulik and Kulik (1992) also provide a brief review.

There are two primary approaches to meta-analysis that are utilized. First, a comparison study analyzes the mean difference hypothesis to determine which of two independent groups or methods scores higher on the variable of interest. Second, the correlation approach investigates relationships between variables under examination. The focus of the present paper is on the process of doing a meta-analysis from the mean difference (comparison) approach.

Steps to Meta-Analytic Research

Before doing a meta-analysis, an exhaustive search to locate a large number of studies on which to base cumulative results must be conducted. This can be done through databases such as ERIC and PsycINFO. Studies of convenience can also be utilized as long as studies are selected randomly from a sample of existing studies (Hunter & Schmidt, 1990).

McNamara (in press) discusses the steps to performing a meta-analysis. The first thing that must be decided is what type of meta-analysis to perform, comparison or correlation. Next, a collective research hypothesis must be determined. Then the meta-analysis population needs to be described and the target population must be determined. An example target population would be patients with Attention Deficit Hyperactivity Disorder (ADHD) who are being treated with a stimulant medication. The collective hypothesis could be that the stimulant-medicated participants' attention would improve more than the non-medicated participants' attention. The meta-analysis population would be people with ADHD treated with stimulant medications for attention. When targeting a population, age limits, geographic locations, treatment conditions, or dosage amounts can aid in specifying the population. For example the meta-analyst may want to

focus on the effects of .3mg/kg of Methylphenidate (Ritalin) on children between the ages six through eighteen with ADHD and no comorbid diagnoses.

The next step suggested is to describe the treatments, interventions or instructional strategies applied across studies. Outcome measures used in the studies then need to be described and categorized. For instance, when measuring stimulant drug effects on attention the researcher would want to look to see what measures were used to measure attention. For example, if Continuous Performance Tests (CPT) were used, and if so, what model, or if parent/teacher ratings were used, and if so, which ones. As part of the meta-analysis what measures are used in each study are coded, so that later the meta-analyst can investigate whether the measures used predict variations in intervention effects (i.e., whether intervention effects are partly measurement artifacts).

The next question to ask is what kind of research designs were used to reach the outcomes the previous studies obtained. For the example using medications, did the studies use controls and double-blind experiments, and how often was medication given, by whom and how much. All these factors can later be important in determining moderator variables (i.e., in predicting variance in the intervention effects across the studies).

Estimating Individual Effect Sizes, *d*

The typical way of reporting meta-analytic results is by an index of effect size. For example, an effect size, *d*, may be estimated for each individual study included in the meta-analysis. As noted by Snyder and Lawson (1993), effect sizes or magnitude of effects inform readers how much of the criterion variable can be "controlled, predicted, or explained" (p. 335). When means and standard deviations are provided, effect sizes are

easily calculated (Kavale, 1982). If a commensurate estimate can in no way be estimated, because the researcher provided insufficient information, the study must be eliminated from the meta-analysis. At this point the meta-analyst then cumulates the *d*-values from many studies.

Researchers are "encouraged" to report their effect sizes (American Psychological Association, 1994, p. 18). Doing so does increase the likelihood that the research will be utilized in future meta-analytic studies (Thompsom, 1997). For instance, consider Kavale's (1982) meta-analysis on the effects of stimulant drugs on hyperactivity. Requirements to be included in their meta-analysis were that studies had to have a comparison group and had to report results in a way that was appropriate for meta-analysis calculations. Out of 500 studies, only 135 met these requirements! Only a few of the studies under review provided means and standard deviations, which is needed for calculating effect size easily. For 104 of the studies, complex solutions from *t* and *F* ratios (Glass, 1977) had to be used to more crudely estimate effect sizes.

For a two-group design (control and experimental group), one effect size is obtained by dividing the sample mean difference by the pooled standard deviation (Glass, 1977, McNamara, 1997). For this example we will look at the control, or unmedicated group, versus the medicated group to find an effect size *g*. Y_m and Y_c are the respective sample means. Hedges and Becker (1986) give the formula for *g* as

$$g = \frac{Y_m - Y_c}{SDp} \qquad (1)$$

The two sample standard deviations (*SD*) and sample sizes are used to find the pooled standard deviation (*SDp*) (Hedges & Becker, 1986). The n_x is the sample size for that group, determined as:

$$SDp = \frac{(n_m-1)SD_{2m} + (n_c-1)SD_{2c}}{n_m + n_c - 2} \quad (2)$$

If a study reports only subgroup means for groups of unequal sizes, the subgroup means need to be weighted and averaged to adjust for differences for group sizes (Hedges & Becker, 1986):

$$Y_x = \frac{n_{x1}Y_{x1} + n_{x2}Y_{x2} + n_{x3}Y_{x3}}{n_x} \quad (3)$$

The effect size estimate *g* is a slightly biased estimate for the estimated population effect size; it tends to exaggerate effect size for the population especially for small samples. This biased estimate needs to be corrected for to find *d* (Hedges & Becker, 1986):

$$d = c_x g = \frac{c_x(Y_m - Y_c)}{SDp}. \quad (4)$$

The values of c_x can be calculated by the following formula (Hedges & Becker, 1986):

$$c_x = 1 - \frac{3}{4n_m + 4n_c - 9} \qquad (5)$$

Hedges and Becker (1986) describe statistical methods for obtaining *d* from ANOVA statistics, mixed-model ANOVA statistics, and correlations. The *d* statistic measures in standard deviation units how far apart the group means are from each other. Some advantages of *d* compared to the point biserial correlation (*r*), explained by Hedges and Becker (1986), are that *d* is not inflated by sample size. Of course, correlational effect sizes can be corrected for sample size (Snyder & Lawson, 1993). Three characteristics that effect size estimates must have to be useful are that the estimates must represent the same construct, must be independent, and must estimate the same statistical parameter (Hedges & Becker, 1986).

Many studies produce more than one effect size. Gleser and Olkin (1994), discuss two sources of multivariate effect size data. The first source is studies that have multiple treatments and report two or more comparisons for the stated dependent variable. The second source is studies with more than one dependent variable with each having a reported effect size. A common procedure in this case is to consider each effect size as independent from the other effect sizes (Rosenthal & Rubin, 1986). Nevertheless, there are numerous theories on how to approach the issue of multiple effect sizes. Statistical methods for determining estimates of effect sizes and population effect size for multivariate studies are described in detail by Hedges and Becker (1986).

Fixed effect regression model. The fixed effect regression model is the model that is most commonly used in meta-analysis with heterogeneous results. This model compares the difference studies to explain variation between studies (Kalain &

Raudenbush, 1986). This model assumes that each study took samples from the same population. Some criticisms of this model made by Erez, Bloom, and Wells (1996) are that the fixed effects model assumes that sampling error explains most of the observed differences in effect sizes and that there is only one fixed and true population correlation.

Random effects model. The random effects model estimates how the between-study differences affect the relationships under investigation. This model considers the studies to be heterogeneous because researchers, samples, and contexts change across studies (Erez, Bloom, & Wells, 1996).

Multivariate mixed linear model. Kalain and Raudenbush (1996) propose that the multivariate model offers meta-analysis the benefits of (a) incorporating multiple effect sizes per study as outcomes, (b) permiting various studies to have different subsets of effect size, and c) treating effect sizes from individual studies as random realizations from a population of potential effect sizes.

Statistical Testing

The last step in a meta-analysis is performing statistical analyses to look at homogeneity of effect sizes and moderator variables (McNamara, 1997). Moderator variables are aspects of the study that vary from one study to the next. Formulas for meta-analysis would be simplified if artifacts (i.e., man-made imperfections) were homogeneous across studies and if there were no moderator variables. If this were true, and sample sizes, standard deviations, and reliability for the scores on the independent and dependent variables were held constant across studies, doing a meta-analysis would simply consist of simple averaging and variance testing (Hunter & Schmidt, 1990). In

the real world though, few things are constant from one study to the next and therefore many factors need to be considered before further calculations are executed.

There are statistical procedures designed specifically for use in a meta-analysis which allow the reviewer to address research questions as if they actually had the raw data (Hedges & Becker, 1986). These procedures can be calculated with SPSS and SAS packages and are derived from effect size properties. Hedges and Becker (1986) describe the complete breakdown of the statistical methods for meta-analysis. For the statistics in this paper it is assumed that the effect size estimator, d, has been corrected for bias by multiplying it by a constant based on the sample size of the study (refer to equation 4).

Theoretical sampling variance of d and confidence intervals must also be obtained and can be calculated from a single observation. Because the unsystematic variance of estimates of the effect size is proportional to $1/n$, where n is the sample size, and each study probably will have a different sample size, the effect size estimates will have different error variances (Hedges & Becker, 1986). For example, take the sample size of non-medicated, or control participants and the medicated participants to find sampling variance (v), which is determined by sample sizes and d (Hedges & Becker, 1986).

$$v = \frac{n_c+n_m}{n_c n_m} + \frac{d^2}{2(n_c+n_m)} \qquad (6)$$

This allows the researcher to make use of the different d-values' degrees of freedoms to estimate systematic effects and still permits the estimating of unsystematic variance used to build statistical tests (Hedges & Becker, 1986). If sample sizes are the same for each

group, $n_m = n_c = n$, then variance can be calculated according to Hedges and Becker (1986) as

$$v = \frac{2}{n}\,\frac{(1 + d^2)}{8} \qquad (7)$$

Confidence intervals can then be calculated by: d-z(alpha) square root of $v \leq$ population effect size $\leq d + z$(alpha) square root of v, where z (alpha) is the 100alpha percent two-tailed critical value of the standard normal distribution (Hedges & Becker, 1986).

Example. Consider the example with medicated and unmedicated treatment for hyperactivity. There are 30 people in the medicated group and 20 people in the control group. The means for the two groups are $Y_m = 35.5$ and $Y_c = 32.8$. The pooled with-in group standard deviation was calculated to be 4.12. The biased effect size g is calculated as:

$$g = \frac{35.5 - 32.8}{4.12} = .655$$

The correction factor is now calculated in order to find the unbiased estimator of effect size d.

$$c_3 = 1 - \frac{3}{4(30) + 4(20) - 9} = .984, \text{ and}$$

$$d = (.655)(.984) = .645$$

The variance of d can now be calculated as

$$v = \frac{20 + 30}{(20)(30)} + \frac{(.645)_2}{2(20 + 30)} = 4.983 .$$

A 95 percent confidence interval can by calculated around the estimated population effect size, d_{pop} as

645– 1.96 (square root of 4.983) is $\leq d_{pop} \leq$ .645+ 1.96 (square root of 4.983)

or

$-3.73 \leq d_{pop} \leq 5.02$.

Variations in d, Effect Sizes

To better understand the results of a meta-analysis, the differences in effect sizes across studies must be examined. Hyde (1986) suggests two steps in analyzing variation in d in the reviewed studies. The first step is to determine if the total variation of effect size values is attributed to random sampling variation. If the effect sizes are homogeneous then all variation can be explained by random sampling variation and the researcher is done. But if the variation is larger than what would be expected by chance and can not be totally attributed to random sampling variation, the researcher must go on to examine systematic variation due to moderator variables. These moderator variables and effect size statistics can then be coded. Glass (1977) stresses the importance of reporting research features in quantitative terms, such as time, IQ, or dose. Nonordinal

features, such as type of treatment, ethnic group, and outcome type, can be coded as indicator variables

Testing homogeneity of effect size. If the population effect size is common across studies then it is reasonable to unite estimates of effect sizes. A problem arises though when the studies do not have a common population effect size. Combining estimates of effect sizes for these studies would result in deception (Hedges & Becker, 1986). Rosenthal and Rubin (1982) provide a test to determine homogeneity of effect size. If the studies examined do not share the same population effect size the homogeneity of effect size will have to be rejected.

Testing for variation in effect size (H_B). When there is a treatment effect and there are variations in effect sizes, the reviewer may want to explain these variations by examining the differences in the studies. One approach to this would be to group studies that have similar properties and little variability in the studies' effect sizes. This permits the reviewer to test the statistical significance of the variation between and within effect size groups (Hedges & Becker, 1986). The process of finding statistically significant variability is similar to the analysis of variance (ANOVA) process. The overall homogeneity statistic H_T must be broken into the parts of between-group homogeneity H_B and within-group homogeneity H_W for the analysis of variance for d (Hedges & Becker, 1986).

$$H_B + H_W = H_T \tag{8}$$

The between-group and within-group homogeneity statistics are comparable to the sum of squares used in ANOVA.

The between-group homogeneity statistic specifies the distribution of true effect sizes across all included studies. H_B equals the weighted sum of the squares of the group mean effect size estimates about the overall mean effect size, all which are weighted. H_B is comparable to the F statistic in ANOVA, in that if H_B exceeds the set critical value then the homogeneity of variance must be rejected and the variation between group mean effect sizes is statistically significant at the alpha level (Hedges & Becker, 1986).

The within-group homogeneity statistic is the sum of the homogeneity statistics for each of the groups. Each individual group is treated as if it were a collection of studies, as follows:

$$H_W = H_{W1} + H_{W2} + \ldots\ldots + H_{Wp} \quad (9)$$

H_W looks at the relationship of the estimated effect size from each group to the true effect size; in other words, H_W represents the estimation error for each group (Kalain & Raudenbush, 1996). The total H_W is the overall test of homogeneity of effect size within the groups of studies, but if there is just one effect size estimate for the entire group, H_W will equal zero. Like the between-group homogeneity, the within-group homogeneity is rejected if it exceeds the critical value (Hedges & Becker, 1986).

Study artifacts. The artifact, sampling error, artificially inflates power to get statistical significance. This type of problem is largely dependent on the size of the sample used. A few other artifacts to consider when doing a meta-analysis are random

measurement errors or unreliability of scores, errors in reported data, and a limited range in independent variables and the dependent variable. The complexity of the formula used in meta-analysis depends on how different the artifacts are across studies, and how much variation there is in the actual correlations (Hunter & Schmidt, 1990).

Kavale (1982) performed a meta-analysis on studies that examined the effects of stimulant medications as a treatment for hyperactivity. Effect sizes and standard deviations were calculated for each outcome class. When attempting to explain variations in effect size measurements, Kavale identified three features that influenced findings. For this meta-analysis the three primary variables identified were study variables, subject variables, and design variables. Study variables included where measures were obtained and by whom. Subject variables consisted of the diagnostic category of hyperactivity in the child. Design variables looked at how the subjects were assigned to groups and the types of control used. It is obvious that these different variables would affect the effect size estimates and need to be investigated.

Meta-analysis is summary an important form of research. With meta-analysis researchers are able to review multiple sources to find a synthesized result and to make sense of the massive amount of information that is available across myriad studies. It is a review of the research that is available. By using the effect sizes from numerous studies that share a hypothesis, a researcher can make a reasonable generalization in regards to the relation between variables. With advances in this form of statistical analysis, researchers will be better able to see more of the whole research picture.

References

American Psychological Association. (1994). Publication manual of the American Psychological Association (4th ed.). Washington, DC: Author.

Erez, A., Bloom, M. C., & Wells, M. T. (1996). Using random rather than fixed effects models in meta-analysis: Implications for situational specificity and validity generalization. Personnel Psychology, 49, 275-306.

Carver, R. (1978). The case against statistical significance testing. Harvard Educational Review, 48, 378-399.

Cohen, J. (1994). The earth is round ($p < .05$). American Psychologist, 49, 997-1003.

Daniel, L. G. (in press). Statistical significance testing: A historical overview of misuse and misinterpretation with implications for the editorial policies of educational journals. Research in the Schools.

Glass, G. V. (1977). Integrated findings: The meta-analysis of research. In L. S. Shulman & S. Mowry (Eds.), Review of Research in Education, 5, 351-379.

Kalain, H. A., & Raudenbush, S. W. (1996). A multivariate mixed linear model for meta-analysis. Psychological Methods, 1, 227-235.

Kavale, K. (1982). The efficacy of stimulant drug treatment for hyperactivity: A meta-analysis. Journal of Learning Disabilities, 15, 280-289.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. Educational and Psychological Measurement, 56,

Hedges, L. V., & Becker, B. J. (1986). Statistical methods in the meta-analysis of research on gender difference. In S. H. Hyde & M. C. Linn (Eds.), The psychology of

gender: Advances through meta-analysis (pp.14-50). Baltimore, MD: John Hopkins University Press.

Hunter, J. E., & Schmidt, F. L. (1990). Methods of meta-analysis: Correcting error and bias in research findings. Sage Publications: Newbury Park, CA.

Hyde, J. S. (1986). Introduction: Meta-analysis and the Psychology of Gender. In S. H. Hyde & M. C. Linn (Eds.),The Psychology of Gender: Advances through meta-analysis (pp.1-13). Baltimore, MD: John Hopkins University Press.

Kulik, J. & Kulik, C. (1992). Meta-analysis: Historical origins and contemporary practice. In B. Thompson (Ed.), Advances in social science methodology (Vol. 2, pp. 53-80). Greenwich, CT: JAI Press.

McLean, J. E., & Ernest, J. M. (in press). The role of statistical significance testing in educational research. Research in the Schools.

McNamara, J. F. (1997). Meta-Analysis: A strategy to improve integrative research reviews. International Journal of Educational Reform, 6, 490-495.

McNamara, J. F. (in press). Conducting your first meta-analysis: An illustrated guide. International Journal of Education Reform.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. Journal of Consulting and Clinical Psychology, 46, 806-834.

Nix, T. W., & Barnette, J. J. (in press). The data analysis dilemma: ban or abandon. A review of null hypothesis significance testing. Research in the Schools.

Rosenthal, R., & Rubin, D. B. (1982). Comparing results of independent studies. Psychological Bulletin, 85, 500-504.

Rosenthal, R., & Rubin, D. B. (1986). Meta-analytic procedures for combining studies with multiple effect sizes. Psychological Bulletin, 99, 400-406.

Schimdt, F.L., & Hunter, J.E. (1997). Development of a general solution to the problem of validity generalization. Journal of Applied Psychology, 62, 529-540.

Synder, P. & Lawson, S. (1993). Evaluating results using corrected and uncorrected effect size estimates. Journal of Experimental Education, 61, 334-349.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. Educational Researcher, 25, 26-30.

Thompson, B. (1997). Computing effect sizes. http://acs.tamu.edu/~bbt0147/

Thompson, B. (1998a, April). Five methodology errors in educational research: The pantheon of statistical significance and other faux pas. Invited address presented at the annual meeting of the American Educational Research Association, San Diego.

Thompson, B. (1998b). In praise of brilliance: Where that praise really belongs. American Psychologist, 53, 799-800.

Thompson, B. (1998c). Review of *What if there were no significance tests?* by L. Harlow, S. Mulaik & J. Steiger (Eds.). Educational and Psychological Measurement, 58, 332-344.

Thompson, B. (in press-a). If statistical significance tests are broken/misused, what practices should supplement or replace them?. Theory & Psychology. [Invited address presented at the 1997 annual meeting of the American Psychological Association, Chicago.]

Thompson, B. (in press-b). Journal editorial policies regarding statistical significance tests: Heat is to fire as p is to importance. Educational Psychology Review.

Thompson, B. (in press-c). Statistical significance tests, effect size reporting, and the vain pursuit of pseudo-objectivity. Theory & Psychology.

Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. Educational and Psychological Measurement, 58, 6-20.